



Skewer: A fast and accurate adapter trimmer for paired-end reads

USER'S MANUAL

Hongshan Jiang

Chinese Academy of Inspection and Quarantine

MAY 12, 2015

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

©Copyright 2013-2015, Chinese Academy of Inspection and Quarantine.
All rights reserved.

Release #	Date	Description
Rev. 0.1.76	09/26/2013	First public release
Rev. 0.1.88	10/16/2013	Barcode demultiplexing
Rev. 0.1.104	01/14/2014	O(kn) worst time complexity
Rev. 0.1.112	03/05/2014	Nextera Long Mate Pair (LMP) adapter trimming; trimming modes (HEAD, ANY, TAIL)
Rev. 0.1.120	09/27/2014	Amplicon paired reads (Bidirectional 5' trimming)
Rev. 0.1.124	03/09/2015	First OS X version
Rev. 0.1.125	05/12/2015	Capability of producing QIIME-compatible files for microbial analysis

Table of Contents

1	General Information	3
1.1	Overview	3
1.2	Citation	3
2	Getting Started	4
2.1	Installation	4
2.1.1	System Requirements	4
2.1.2	Install from Binary Package	4
2.2	For Impatient Users	5
3	Using the Program	7
3.1	Usage	7
3.2	Options	7
3.2.1	Adapter	7
3.2.2	Tolerance	9
3.2.3	Filtering	9
3.2.4	Input/Output	10
3.2.5	Miscellaneous	11
4	Source Codes	12

Chapter 1

General Information

1.1 Overview

The *Skewer* program implements a novel dynamic programming algorithm dedicated to the task of adapter trimming and it is specially designed for processing Illumina® paired-end sequences. *Skewer* got its name because the implemented algorithm utilizes the equality of diagonal adjacent elements in the dynamic programming matrix where these elements could be ‘skewered’ together by such a tool.

Skewer has the following features:

- Detection and removal of adapter sequences
- Insertion and deletion allowed in pattern matching
- Targeted at Single End, Paired End (PE), and Long Mate Pair (LMP) reads
- Demultiplexing of barcoded sequencing runs
- Multi-threading support
- Trimming based on phred quality scores
- IUPAC characters for barcodes and adapters
- Compressed input and output support

The program is available at <https://sourceforge.net/projects/skewer>.

1.2 Citation

Jiang, H., Lei, R., Ding, S.W. and Zhu, S. (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, **15**, 182.

Chapter 2

Getting Started

2.1 Installation

2.1.1 System Requirements

CPU	64 bit Intel or AMD CPU
RAM	30M per CPU (core)
Operating System	x86_64 GNU Linux
Pre-requisites	none

2.1.2 Install from Binary Package

Instructions: Move skewer-x.x.x-linux-x86_64 to a directory 'INSTALL_DIR' where the software is to be installed.

```
$ mv skewer-x.x.x-linux-x86_64 INSTALL_DIR/skewer
```

For convenience, you should add 'INSTALL_DIR/' into the environmental variable \$PATH by adding a line 'export PATH=INSTALL_DIR:\$PATH' to ~/.bashrc followed by command '. ~/.bashrc'.

```
$ echo "export PATH=INSTALL_DIR:$PATH" >> ~/.bashrc
$ . ~/.bashrc
```

2.2 For Impatient Users

Basically, the usage format of *skewer* is:

```
$ skewer [options] <reads-pair1.fastq> [<reads-pair2.fastq>]
```

where the primary options are as follows:

- options ‘-x’ and ‘-y’ are required for most of the cases, for specifying the forward adapter(s)/primer(s) and the reverse adapter(s)/primer(s);
- option ‘-m’ is used for specifying trimming mode, such as *head* mode, *tail* mode, *anywhere* mode, *paired-end* mode, *mate-pair* mode, and *amplicon* mode;
- options ‘-r’, ‘-d’, and ‘-k’ are used for specifying trimming stringency;
- options ‘-q’ and ‘-Q’ are used for quality-based trimming or filtering;
- options ‘-l’ and ‘-L’ are used for length-based filtering;
- option ‘-t’ is used for multithreading.

Examples are as follows:

1. To trim the sequences in *adapters.fa* from *sample.fastq*, filter out those reads that have an average phred-score below 9, and output the trimmed reads to *out-trimmed.fastq*:

```
$ skewer -Q 9 -t 2 -x adapters.fa sample.fastq -o out
```

2. To do adapter trimming for the compressed files *data-pair1.fq.gz* and *data-pair2.fq.gz* with the adapter sequence of AGATCGGAAGAGC, and do 3’ end quality trimming with a quality threshold of 3:

```
$ skewer -x AGATCGGAAGAGC -q 3 data-pair1.fq.gz data-pair2.fq.gz
```

3. To trim adapter sequence TCGTATGCCGTCTTCTGCTTGT from small-RNA reads *srna.fastq*, not allowing insertion/deletion in adapter detection, and only output those reads that have a length between 16 and 30 after adapter trimming:

```
$ skewer -x TCGTATGCCGTCTTCTGCTTGT -l 16 -L 30 -d 0 srna.fastq
```

4. To do adapter trimming from *lmp-pair1.fastq* and *lmp-pair2.fastq* using Long Mate Pair mode, and redistribute reads based on junction information:

```
$ skewer -m mp -i lmp-pair1.fastq lmp-pair2.fastq
```

5. To trim barcodes which are the 6 leading nt at the 5' end of the reverse primer of the amplicon sequences in *mix-pair1.fastq* and *mix-pair2.fastq*, output those barcodes to *barcodes.fastq* and *mapping-file.txt* as well as the trimmed reads for downstream analysis (such as those by QIIME):

```
$ skewer -m ap --cut 0,6 --qiime -x forward-primers.fa \
-y reverse-primers.fa mix-pair1.fastq mix-pair2.fastq
```

where an example of the *mapping-file.txt* required by QIIME looks like this:

```
#SampleID BarcodeSequence LinkerPrimerSequence ReversePrimer Description
A01 TCGAAT GTGCCAGCMGCCGCGGTAA GGACTACHVGGGTWTCTAAT NA
A02 CATGGC GTGCCAGCMGCCGCGGTAA GGACTACHVGGGTWTCTAAT NA
A03 AGCTTA GTGCCAGCMGCCGCGGTAA GGACTACHVGGGTWTCTAAT NA
A04 GTACCG GTGCCAGCMGCCGCGGTAA GGACTACHVGGGTWTCTAAT NA
A05 CTATGA GTGCCAGCMGCCGCGGTAA GGACTACHVGGGTWTCTAAT NA
A06 ACGCTG GTGCCAGCMGCCGCGGTAA GGACTACHVGGGTWTCTAAT NA
A07 GATACT GTGCCAGCMGCCGCGGTAA GGACTACHVGGGTWTCTAAT NA
A08 TGCGAC GTGCCAGCMGCCGCGGTAA GGACTACHVGGGTWTCTAAT NA
A09 GCTTAA GTGCCAGCMGCCGCGGTAA GGACTACHVGGGTWTCTAAT NA
A10 TACCGG GTGCCAGCMGCCGCGGTAA GGACTACHVGGGTWTCTAAT NA
A11 CGAATT GTGCCAGCMGCCGCGGTAA GGACTACHVGGGTWTCTAAT NA
A12 ATGGCC GTGCCAGCMGCCGCGGTAA GGACTACHVGGGTWTCTAAT NA
```

Chapter 3

Using the Program

3.1 Usage

```
$ skewer [options] <reads-pair1.fastq> [<reads-pair2.fastq>]
```

3.2 Options

3.2.1 Adapter

```
-x <string>
```

Adapter sequence/file for the first reads. If it's not specified, the adapter sequence is 'AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC' by default. If there's a dot in the string, then it's recognized as the filename of the FASTA file that contains adapter sequences.

```
-y <string>
```

Adapter sequence/file for the second reads. If it's not specified, the adapter sequence is 'AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA' by default. If there's a dot in the string, then it's recognized as the filename of the FASTA file that contains adapter sequences. If -x is the only one specified explicitly, then -y is implied by -x.

```
-M, --matrix <string>
```

TAB delimited file indicates valid forward/reverse adapter pairing. For the specified forward and reverse adapter sequences, the matrix contained in the

specified file further denotes the valid combinations. It is an all ones matrix by default.

The following matrix specifies that the forward/reverse adapter combinations A2, A3, B2, B3, C1, D1, D2, D3, and D4 are valid, where forward adapters are numbered as 'A', 'B', 'C', 'D', ... while reverse adapters are numbered as '1', '2', '3', '4', ... `%*` or `*%` denote match of a single adapter.

#	%	1	2	3	4
%	0	0	0	0	0
A	0	0	1	1	0
B	0	0	1	1	0
C	0	1	0	0	0
D	0	1	1	1	1

`-j <string>`

Junction adapter sequence/file for Nextera Mate Pair reads. The junction adapter sequence is 'CTGTCTCTTATACACATCTAGATGTGTATAAGAGACAG' by default. If there's a dot in the string, then it's recognized as the filename of the FASTA file that contains junction adapter sequences.

`-m, --mode <string>`

trimming mode. For single end reads, the valid modes are '*head*' for 5' end trimming; '*tail*' for 3' end trimming; '*any*' for anywhere adapter detection and trimming. For paired end reads, the valid modes are '*pe*' for paired-end trimming; '*mp*' for mate-pair trimming; '*ap*' for amplicon trimming besides '*head*', '*tail*', and '*any*' which specifies separate single end trimming for paired reads. In amplicon mode, the forward/reverse primers will not be trimmed off because they are informative for downstream analysis.

`-b, --barcode`

Whether to demultiplex reads according to adapters/primers. The default is *no*.

`-c, --cut <integer>,<integer>`

To hard clip off the 5' leading bases of the forward primer and reverse primer respectively as the barcodes in amplicon mode. The default is *0,0*.

3.2.2 Tolerance

`-r <number>`

Maximum allowed error rate. The error rate is defined as the normalized number of errors (refer to section 1.2) considering the phred quality scores divided by the length of aligned region. The valid range of error rate is $[0, 0.5]$. The default value is 0.1.

`-d <number>`

Maximum allowed indel error rate. The indel error rate is defined as the number of indels multiplied by the maximum penalty of a mismatch (refer to section 1.2) divided by the length of aligned region. The valid range of indel error rate is $[0, r]$, where r is the maximum allowed error rate. The default value is 0.03.

`-k <integer>`

Minimum overlap length for adapter detection. For single-end reads trimming or single-end trimming of paired-end reads, the default value is $\max(1, \text{int}(4 - 10 * r))$. For mate-pair reads, the default value is half of the length of corresponding junction adapter.

3.2.3 Filtering

`-q, --end-quality <integer>`

3' end quality trimming. Trim 3' end until specified or higher quality reached. The default value is 0.

`-Q, --mean-quality <integer>`

Reads filtering by average quality. Specifies the lowest mean quality value allowed before trimming. The default value is 0.

`-l, --min <integer>`

Minimum read length allowed after trimming. The default value is 18.

`-L, --max <integer>`

Maximum read length allowed after trimming. There's no limit by default.

-n

Whether to filter out highly degenerative reads, where highly degenerative reads are defined as those reads that above 15% of the nucleotides are 'N'. The default is *no*.

-u

Whether to filter out undetermined mate-pair reads, where undetermined mate-pair reads are defined as those mate-pair reads that neither of the paired reads has the junction adapter sequence detected. The default is *no*.

3.2.4 Input/Output

-f, --format <string>

Format of FASTQ quality value. The valid options are 'sanger', 'solexa', 'auto'. The default value is 'auto'.

-o, --output <string>

Base name of output file. This specifies the prefix of output files. The default value is '<reads>-trimmed', where <reads> is the base name of input file.

-z, --compress

Whether to compress output in GZIP format. The default is *no*.

-1, --stdout

Whether to redirect output to STDOUT. This option suppresses the '-b', '-o', and '-z' options. The default is *no*.

--qiime

Whether to prepare files required by QIIME. If specified, the "barcodes.fastq" and "mapping_file.txt" will be prepared for downstream analysis with QIIME. This option must be used with '-cut' option and will suppress '-barcode' option. The default is *no*.

--quiet

Whether in quiet mode. When specified, there will be no progress update. The default is *no*.

3.2.5 Miscellaneous

`-i, --intelligent`

Whether to intelligently redistribute reads. When specified, mate-pair reads will be redistributed based on detected junction information. The default is *no*.

`-t, --threads <integer>`

Number of concurrent threads. The valid number is an integer between 1 and 32. The default value is 1.

Chapter 4

Source Codes

Source codes are deposited at <https://github.com/relipmoc/skewer>.